Abstract

Effect of Features Generated from Adjacent and Overlapped Segments in Protein Sequence Classification

Graduate School of

Natural Science & Technology

Kanazawa University

Division of Electrical Engineering and Computer Science

Student ID No. 1524042012

Name: Mohammad Reza Faisal

Chief Advisor: Professor Kenji Satou

Date of Submission: 29 June 2018

Abstract

In protein sequence classification research, sequences must be converted into data that are understood by classification algorithms. Protein descriptor is the name of the tool to convert sequence into feature representation. There is two type of protein descriptor: the first is alignment-based descriptor or position-specific descriptor. The second is a position-independent descriptor.

Position-independent descriptors convert a variable length sequence of protein into fixed length numerical features. These descriptors are useful since they apply to any length of a sequence, however, positional information of subsequence is discarded even though it might have a high contribution to classification performance. To solve this problem, we divided the original sequence into some segments. We generated to kind of segments those are adjacent segments and overlapped segments. Then we calculated the numerical features for them.

Features generated from adjacent and overlapped segments enables us to partially introduce positional information (for instance, compositions of serine in anterior and posterior segments of a sequence). Through comprehensive experiments on the number of segments and length of the overlapping region, we found our classification approach with sequence segmentation and feature selection is effective to improve the performance. We evaluated our approach on three protein classification problems, i.e., classification of nuclear receptors, protein family classification, and cell-penetrating peptides prediction. We achieved significant improvement in all cases which have a dataset with sufficient amino acid in each sequence. This result has shown the great potential of using additional segments in protein sequence classification to solve other sequence problems in bioinformatics.

Keyword: protein sequence classification, protein descriptor, sequence segmentation, feature selection

1. Introduction

1.1 Background

The protein sequence is an essential asset in protein classification research. To apply different machine learning approaches on protein sequence, it is a standard process to convert protein sequence into a feature representation. This process is called feature extraction, and it is an important step because the choice of the effective type of feature extraction will affect classification performance. It drives the scientists to develop algorithm or method that performs feature extraction process, which is commonly known as protein descriptors.

In last two decades, researchers have developed many protein descriptors. Moreover, those descriptors have been used to solve the various case of protein analysis. From all of those developed descriptors, 22 type descriptors have been actively used in researchers. Those descriptors can be grouped into eight groups such as Amino Acid Composition, Autocorrelation, CTD, Conjoint Triad, Quasi-Sequence-Order, Pseudo-Amino Acid Composition, Proteochemometric descriptors, and Profile-based descriptor.

The profile-based descriptor is alignment-based descriptor or position-specific descriptor that convert a sequence based on the Position Specific Scoring Matrix (PSSM). The feature representation of this descriptor often shows good performance, because it has position information of a sequence. However, the length of feature representation may vary and depend on the length of the protein sequence. Other groups of descriptors are position-independent descriptor or alignment-free descriptor. These descriptors convert a variable length sequence of protein into a fixed length feature representation. These descriptors are useful since they apply to any length of the sequence.

One common thing in these researchers is that only a full length of the sequence is used as an input to the protein descriptor. It means that the output of the protein descriptor only describes the state of a whole protein alone. In the use of position-specific descriptor, generated feature representation from only a full length of the sequence may enough because there is position information in that feature representation. However, the length of feature representation may vary, and it depends on the number of amino acid in a sequence. The variation of the length of feature representation makes it difficult to use on classification algorithms since they require the same number of feature representation. Because of that, it is popular to convert a variable length sequence of protein into a fixed length feature representation by using position-independent descriptors. However, positional information of subsequence is discarded even though it might have a high contribution to classification performance.

2. Literature Review

2.1 Classification of Nuclear Receptors

Nuclear receptors are key transcription factors that manage important gene networks responsible for cell growth, differentiation, and homeostasis [1]. Classification of nuclear receptors was done in researches [1],[2].

As done by Bhasin and Gajendra [1], the classification was achieved by amino acid composition and dipeptide composition from a sequence of nuclear receptors using support vector machine (SVM). The performance of both classifiers was evaluated using 5-fold cross-validation. The accuracy of the amino acid composition-based classifier was 82%, and dipeptide composition-based classifier was 97.5%.

In the research done by Wang et al. [2], the classification was achieved by various protein descriptors from a sequence of nuclear receptors using Fuzzy K nearest neighbor (FK-NN). They create two layers of the predictor. The first layer was used to identify a query protein as NR or not. In the second layer was used to identify the NR among the seven subfamilies. The performance of all classifier was evaluated using jackknife test and independent dataset test. The overall accuracy of first layer predictor is 92.56% by using jackknife test and 98.03% by using independent dataset test. Moreover, the overall accuracy of second layer predictor is 88.68% by using jackknife test and 99.65% by using independent dataset test.

2.2 Protein Family Classification

A protein family is a set of proteins that are evolutionarily related, typically involving similar structures or functions [3]. Protein family classification was done in researches [3], [4]. Cai et al. [4] had classified 54 functional families. The feature extraction process had been done by using a combination of protein descriptors which are composition, translation, and distribution. The reported accuracies of family classification had been in the range of 69.1 - 99.6%. In another study, Asgari and Mofrad [3] performed classifications of 7,027 protein families. They applied a new feature extraction method as known as ProtVec. The average accuracy for the first 1000 families was obtained 94% ± 0.05% by using SVM with 10-fold cross-validation.

2.3 Cell-Penetrating Peptides Prediction

Cell-penetrating peptides (CPPs) are small peptides that are about 10–30 amino acids long. CPPs can carry various bioactive cargoes, ranging from small molecules to proteins and supramolecular particles, to directly enter cells without significantly damaging the cell membrane. It makes them potential drug delivery agents for the translocation of cargo into cells. CPP prediction research has increased in the past few years. CPPsite2.0 is CPP-specific database that has approximately 1850 experimentally validated CPPs [5].

CPPred-RF is one method that has succeeded to solve the CPPs prediction case [5]. In this study Wei et al. used two datasets that are CPP924 and CPPsite3. In feature extraction process, they used parallel correlation pseudo-amino-acid composition (PC-PseAAC), series correlation pseudo-amino acid composition (SC-PseAAC), adaptive skip dipeptide composition (ASDC) and physicochemical properties (PPs). The result is numerical representation with 636 features. Then features selection is applied by using Max-Relevance-Max-Distance (MRMD) as feature ranking method and Sequential Feature Selection (SFS) as optimal features selector. Moreover, they used the random forest as the classifier with jackknife test at the prediction and evaluation stage. The result is 91.6% Accuracy for CPP924 dataset and 71.1% CPPsite3.

2.4 Implementation of Existing Protein Descriptor

R package protr has various structures and physicochemical descriptors and PCMs modeling descriptors for amino acid sequence [6]. protr has eight group descriptors. The first seven groups are the alignment-free descriptors and the last group, PSSM, is an alignment-based descriptor. The PSSM group has PSSM profile descriptor that produces outputs with a varying number of features depends on the number of amino acid.

In active research on protein classification, feature extraction is one of the important processes. This process converts a protein sequence into numerical features by using protein descriptor. The protein descriptor can then be written as the following formula:

$$descriptor(s) = f \tag{1}$$

The output of $descriptor(s)$ is numerical features f.

The use of a single protein descriptor based classifier has solved protein analysis cases. It predicts nuclear receptor [1], membrane protein types [7], protein folding [8], protein-protein interaction (PPI) [9], and protein subcellular locations [10]. It also detects the remote homology and folds recognition [11]. A combination of various descriptors is also used to generate a numerical representation of protein sequence in general active research. This formula can represent a combination of various descriptors implementation:

$$\bigcup_{type} descriptor_{type}(s) = \bigcup_{type} f_{type} \tag{2}$$

Where *type* is descriptor type, *type* ∈ {amino acid composition, dipeptide composition, tripeptide composition, and other descriptors}.

One of the successful reports of this approach is the study of predicting protein functional families by using a combination of eight descriptors from alignment-free groups [12]. Moreover, the other study used a combination of alignment-free descriptors and alignment-based descriptors for remote protein

homology detection [13]. Both of that studies had same conclusion that the combination of various descriptors can give a better result than using a single descriptor only.

3. Data and Methods

3.1 Dataset

We used datasets from three protein analysis cases in this research:

1. Classification of Nuclear Receptors.

   This dataset was used in Wang et al. research [2]. They used 159 sequences of nuclear receptors obtained from NucleaRDB and 500 sequences of non-nuclear receptors obtained from UniProt database. Detail dataset description is shown in Table 1.

   Table 1. The description of the dataset in Wang et al. research.

No	Set	Subfamily	# sequence
1	Nuclear receptors (NR)	NR1: thyroid hormone-like	50
2		NR2: HNF4-like	36
3		NR3: estrogen-like	37
4		NR4: nerve growth factor IB-like	7
5		NR5: Fushi tarazu-F1 like	12
6		NR6: germ cell nuclear factor like	5
7		NR0: knirps and DAX like	12
8	Non-nuclear receptors (Non-NR)	N/A	500

2. Protein Family Classification.

   Protein family dataset was used in Asgari and Mofrad research [3]. They obtained the dataset from Swiss-Prot database. The dataset has 7,027 protein families of 324,018 protein sequences. We only used 1,000 protein families in our research.

3. Cell-Penetrating Peptides Prediction.

   Wei *et al.* used two datasets in cell-penetrating peptides prediction research [5]. Datasets were obtained from CPPsite2.0. Detail dataset description is shown in Table 2.

   Table 2. Dataset Description of the dataset in research [5].

No	Dataset	# positive	# negative	# amino acid
1	CPP924	462	462	10 – 61
2	CPPsite3	187	187	5 – 61

3.2 Methods

3.2.1 Flowchart of Research Method

Our proposed approach consists of main three steps. The flowchart of our approach is explained in Figure 1.

The first step is feature extraction that has three processes:

1. Sanity check of the amino acid types is responsible for erasing amino acids if they are not in the 20 default of amino acid types.
2. Sequence segmentation is conducted for dividing a sequence into adjacent segments and overlapped segments.
3. Feature construction is in charge of converting an original sequence, adjacent segments, and overlapped segments into numerical features by using existing descriptor from protr package. Then a concatenation of all those numerical features is created.

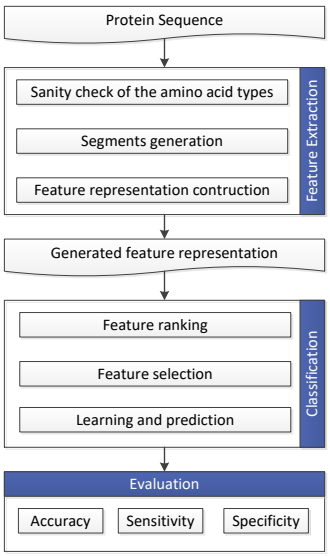

Figure 1. Research method flowchart.

The second step is classification. We conduct k-fold cross-validation or jackknife test, each process in this step are repeated k times or n time, with n is a number of samples.

1. Feature ranking is responsible for sorting features by importance. The random Forest function for R [14] conducts this process.
2. Feature selection and prediction are responsible for creating feature subsets, and performing learning and predicting with ksvm function in a kernlab package for R [15].

The last step is the evaluation. It is in charge of calculating accuracy for prediction result. We also investigated the important features in feature subset which gave the best classification performance.

3.2.2 Segments Generation

We show Equations (1) and (2) that can represent the feature extraction process that has been used in active research. One common thing in both equations is that they use a full-length of sequence s as the input. Moreover, f is the output which provides global information of s.

Our approach generates segments as additional input. There are two type of segments namely adjacent segment and overlapped segment. The adjacent segment is generated from the first segment is calculated from the beginning of the sequence, then followed by the second segment and so on. The

overlapped segment is generated by merging the half from the end of the first segment and a half from the beginning of the second segment.

3.2.3 Feature Representation Construction

After segments are created, we calculate features of sequence *s* by using the formula below:

$$descriptor(s) \cup \bigcup_{k=2}^{z} \left(\left(\bigcup_{i=1}^{k} descriptor(segment_i) \right) \cup \left(\bigcup_{l=1}^{k-1} descriptor(overlapped_l) \right) \right) \quad (3)$$

We can also implement this approach with a combination of various descriptors by using the formula below:

$$\bigcup_{type} \left(descriptor_{type}(s) \cup \bigcup_{k=2}^{z} \left(\left(\bigcup_{i=1}^{k} descriptor_{type}(segment_i) \right) \cup \left(\bigcup_{l=1}^{k-1} descriptor_{type}(overlapped_l) \right) \right) \right) \quad (4)$$

4. Results and Discussion

We compare our result with the result from previous researches. We also show the investigation result on important features of the feature subset which give the best performance.

4.1 Dataset of Classification of Nuclear Receptors

In this protein classification case, we conducted two experiments. In the first experiment, we compared our approach result with experiment result from Bhasin and Gajendra [1]. In this experiment, we converted a sequence into features representation by using Eq. 3. We generated two type of feature representation. The first feature representation was generated by using AAC, and the second is generated by using DC. We conducted classification process by using SVM with 5-fold cross-validation test. The result comparison is shown in Table 3.

Table 3. The result comparison of our approach and method in research [1].

No	Method	Accuracy (%)	# Features	Description
1	AAC	67.99	20	AAC based classifier of Research [1].
2	DC	93.60	400	DC based classifier of Research [1].
3	AAC_7	86.97	980	AAC based classifier with z = 7.
4	DC_4	94.19	6400	DC based classifier with z = 4.
5	AAC_7 FS	88.06	790	AAC based classifier with z = 7 and feature selection.
6	DC_4 FS	**96.19**	355	DC based classifier with z = 4 and feature selection.

The second experiment performed to compare our approach with research of Wang et al. [2]. The result for identifying NR and non-NR is shown in Table 4.

Table 4. Detail comparison of our approach and method in research [2] for identifying NR and non-NR.

No	Method	Accuracy (%)	# Features	Description
1	NR-2L	92.56	881	Result by Wang *et al*.
2	AAC_3	97.56	180	AAC based classifier with z = 3

3	DC_2	97.87	1600	DC based classifier with z = 2
4	AAC_3 FS	97.87	100	AAC based classifier with z = 3 and feature selection
5	DC_2 FS	**98.48**	120	DC based classifier with z = 2 and feature selection

In the second level experiment, we identified NR subfamilies. The detail comparison result is shown in Table 5.

Table 5. Detail comparison of our approach and method in research [2] **for identifying NR subfamilies.**

No	Method	Accuracy (%)	# Features	Description
1	NR-2L	88.68	881	Result by Wang *et al*.
2	AAC_5	81.76	500	AAC based classifier with z = 5
3	DC_2	91.81	1600	DC based classifier with z = 2
4	AAC_5 FS	83.01	355	AAC based classifier with z = 5 and feature selection
5	DC_2 FS	**94.33**	145	DC based classifier with z = 2 and feature selection

In this protein classification case, we showed our approach could work better than two previous researches. As we can see in Table 3, we also succeed to reduce features by using feature ranking and feature selection. We reduced feature of generated AAC feature representation from 980 to 790 features and generated DC feature representation from 6400 to 355 features. Moreover, the detail of important features of generated DC feature representation with z=4 is shown in Table 6.

Table 6. Detail of important features in DC_4 FS experiment.

Source	# Important Feature	# Features before feature selection
Original sequence	34	400
k = 2	90	1200
k = 3	124	2000
k = 4	107	2800
Total	355	6400

In the second experiment, we also showed our approach has better performance than NR-2L method [2]. In a comparison of NR and non-NR prediction performance, the best performance was obtained when implementing our approach on DC based classifier with feature selection. Feature selection process reduced features of generated AAC feature representation from 180 to 100 features and generated DC feature representation was reduced from 1600 to 120 features. Detail of important features of DC_2 FS are shown in Table 7.

Table 7. Detail of important features in generated DC feature representation with z=2.

Source	# Important Feature	# Features before feature selection
Original sequence	37	400
k = 2	83	1200
Total	120	1600

In a comparison of identifying NR subfamilies, the high improvement was obtained when implementing our approach on DC based classifier. Moreover, the detail of important features of DC_2 FS experiments are shown in Table 8.

Table 8. Detail of important features of DC_2 FS experiment.

Source	# Important Feature	# Features before feature selection
Original sequence	43	400
k = 2	102	1200
Total	145	1600

4.2 Dataset of Protein Family Classification

In this experiment, we used the dataset that was provided by Asgari and Mofrad [3] and performed 1000 classification cases using the first 1000 families. The classification performed in this experiment is a balanced binary classification. Samples of the positive class are samples of a selected protein family. Samples of the negative class are randomly selected samples. In the feature extraction process, we used a combination of various protein descriptors which are Amino Acid Composition (AAC), Composition (CTDC), translation (CTDT), and distribution (CTDD) with z = 5. Moreover, we used SVM with 10-fold cross-validation test as classifier and evaluation method. We calculated weighted accuracy from 1000 classification experiments, and the result is shown in table 9. We found our method has better accuracy than the previous method.

Table 9. Prediction accuracy comparison of our approach and method in research [3] for classifying first 1000 families.

No	Method	Description	Weighted Specificity	Weighted Sensitivity	Weighted Accuracy (%)
1	ProVec 1000	Method from [3]	0.920802	**0.949276**	93.95
2	Our Approach	Our method	0.98791	0.935978	96.19
3	Our Approach FS	Our method with feature selection	**0.989965**	0.947138	**96.79**

We have investigated subset features that can obtain the best accuracy prediction from each family classification case. The result of our investigation of three families, one of the family result is shown in Table 10. We show a subset features were formed of the four descriptors that we used with all various k values.

Table 10. Detail of important features in 50S ribosome-binding GTPase family classification.

protein descriptor	# features from sequence					# total important feature
	original	k = 2	k = 3	k = 4	k = 5	
AAC	13	36	53	64	84	250
CTDC	13	47	87	110	129	386
CTDT	11	35	55	79	98	278
CTDD	76	165	237	295	263	1036

4.3 Dataset of Cell-Penetrating Peptides Prediction

We implemented our approach as single descriptor and combination of various descriptors based classifier. We used AAC, PseAAC, DC and composition/distribution/translation (CTD) descriptor on feature extraction process. In the classification and evaluation process, we used SVM as a classifier

with 10-fold cross-validation test. Our approach cannot give a better performance than CPPred-RF [5]. Feature representation from additional segments made performance decrease in most of all experiment that we did. We assumed it happened because sequences in dataset CPP924 and CPPsite3 did not have sufficient amino acids as we can see in Table 11.

Table 11. Statistic comparison of amino acid numbers in sequences.

No	Protein Classification Case	Number of Amino Acid				
		Min	Max	Median	Mean	Mode
1	Classification of Nuclear Receptor	2	3932	419	510	419
2	Protein Family Classification	7	21531	332	425	101
3	Cell-Penetrating Peptides Prediction	5	61	17	19	18

5. Summary and Future Work

We developed a simple and robust approach for protein sequence classification. We generated a novel feature representation by merging of feature representation of original sequence, adjacent and overlapped segments. We implemented feature ranking and feature selection to reduce the noise and to look for important features. We succeed to improve classifier performance. We showed the best feature subset contains some feature from feature representation that used the various value of divider. It means additional segments contribute to improving classifier performance.

Our approach achieved significant improvement in all cases which have a dataset with sufficient amino acid in each sequence. We evaluated our approach on three protein analysis cases. It worked as a single descriptor and a combination various descriptors based classifier.

Fifteen other alignment-free protein descriptors can be used with our proposed method in the future research. Also, we will implement our approach to solve other sequence problems in bioinformatics, such as DNA sequence classification.

6. References

[1] M. P. S. R. Bhasin and Gajendra, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *J. Biol. Chem.*, vol. 279, no. 22, 2004.

[2] P. Wang, X. Xiao, and K.-C. Chou, "NR-2L: A Two-Level Predictor for Identifying Nuclear Receptor Subfamilies Based on Sequence-Derived Features," *PLoS One*, vol. 6, no. 8, p. e23505, Aug. 2011.

[3] E. Asgari and M. R. K. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PLoS One*, vol. 10, no. 11, pp. 1–11, 2015.

[4] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic*

Acids Res., vol. 31, no. 13, pp. 3692–3697, Jul. 2003.

[5] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou, "CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency," *J. Proteome Res.*, vol. 16, no. 5, pp. 2044–2053, May 2017.

[6] N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu, "protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences," *Bioinformatics*, vol. 31, no. 11, pp. 1857–1859, 2015.

[7] Z.-P. Feng and C.-T. Zhang, "Prediction of Membrane Protein Types Based on the Hydrophobic Index of Amino Acids," *J. Protein Chem.*, vol. 19, no. 4, pp. 269–275, 2000.

[8] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 19, pp. 8700–8704, Sep. 1995.

[9] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein–protein interactions based only on sequences information," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 11, pp. 4337–4341, Mar. 2007.

[10] K.-C. Chou, "Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect," *Biochem. Biophys. Res. Commun.*, vol. 278, no. 2, pp. 477–483, 2000.

[11] H. Rangwala and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition," *Bioinformatics*, vol. 21, no. 23, pp. 4239–4247, Dec. 2005.

[12] S. A. K. Ong, H. H. Lin, Y. Z. Chen, Z. R. Li, and Z. Cao, "Efficacy of different protein descriptors in predicting protein functional families," *BMC Bioinformatics*, vol. 8, p. 300, Aug. 2007.

[13] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation," *Mol. Inform.*, vol. 32, no. 9–10, pp. 775–782, 2013.

[14] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[15] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab -- An {S4} Package for Kernel Methods in {R}," *J. Stat. Softw.*, vol. 11, no. 9, pp. 1–20, 2004.

平成３０年　８月　６日

学位論文審査報告書（甲）

１．学位論文題目（外国語の場合は和訳を付けること。）

Effect of Features Generated from Adjacent and Overlapped Segments in Protein Sequence Classification（隣接し重なり合う部分配列から生成された特徴がアミノ酸配列分類にもたらす効果）

２．論文提出者　(1) 所　　属　電子情報科学　専攻

(2) 氏（ふり）　名（がな）　Mohammad Reza Faisal（もはまど　れざ　ふぁいさる）

３．審査結果の要旨（600～650 字）

平成３０年８月６日に第１回学位論文審査委員会を開催し、同日に口頭発表、その後に第２回審査委員会を開催し、慎重審議の結果、以下の通り判定した。なお、口頭発表における質疑を最終試験に代えるものとした。

DNA 配列やタンパク質のアミノ酸配列を入力として、その配列がある性質を持つかどうかを予測する場合、一般的には可変長配列から k-mer 頻度などの数量的な特徴を計算し、数値ベクトルとしての特徴ベクトルを生成して、分類器に入力する。しかし、このような数量的特徴は一般に、計算の過程で配列上の位置情報が失われるという欠点がある。本研究では入力配列を予め分割し、位置情報をある程度取り込んだ特徴ベクトルを生成することにより、予測精度を向上できることを示した。このような手法は殆ど研究されていないが、分割数を適宜調整して分割部分がオーバーラップするよう分割することにより、データによっては従来手法を超える精度を達成できることを示した。

以上の研究成果は、可変長 DNA 配列の分類問題における精度向上の新たな方向性を示すものであり、本論文は博士（工学）に値するものと判定した。

４．審査結果　(1) 判　　定（いずれかに○印）　（合　格）・　不合格

(2) 授与学位　　博　士（工学）